THE 1997 ABBOT SYSTEM FOR THE TRANSCRIPTION OF BROADCAST NEWS

G.D. Cook *A.J. Robinson*

Cambridge University Engineering Department,
Trumpington Street,
Cambridge, CB2 1PZ, UK

ABSTRACT

This paper describes the development of a connectionist-hidden Markov model (HMM) system for the 1997 DARPA Hub-4E CSR evaluations. We describe both system development and the enhancements designed to improve performance on broadcast news data. Both multilayer perceptron (MLP) and recurrent neural network acoustic models have been investigated. We assess the effect of using gender-dependent acoustic models, and the impact on performance of varying both the number of parameters and the amount of training data used for acoustic modelling. The use of context-dependent phone models is described, and the effect of the number of context classes is investigated. We also describe a method for incorporating syllable boundary information during search. Results are reported on the 1997 DARPA Hub-4E development test set. We then describe the CU-CON evaluation system and report results on the 1997 Hub-4E test set.

1. Introduction

This paper describes experiments aimed at improving the performance of the ABBOT system on broadcast news data. ABBOT is a large vocabulary connectionist-HMM continuous speech recognition system developed at Cambridge University Engineering Department [1]. The connectionist-HMM approach uses an underlying hidden Markov process to model the time-varying nature of the speech signal and a connectionist system to estimate the observation likelihoods within the hidden Markov model framework [2].

The layout of this paper is as follows. We first describe acoustic modelling experiments. The use of MLP acoustic models is described, and results are reported for both gender-independent and gender-dependent systems. Section 2.2 describes the use of recurrent neural network models. We examine the effect of both model size (in terms of the number of model parameters) and the amount of training data on recognition performance. The effect of the number of context-dependent phone models is also investigated.

Section 3 describes a technique for incorporating syllable boundary information during the decoding procedure. We describe the method used to determine the syllable boundary points in the training data, and how this is used to train a syllable onset detector. The syllable boundary information has been incorporated in the decoding procedure without the need to modify the decoder, and the method by which this is achieved is described.

The CU-CON system which participated in the 1997 DARPA Hub-4E evaluation is then described. This includes details of the acoustic features and the acoustic and language models. The two-pass recognition strategy employed by the system is also outlined. Official results on the 1997 evaluation data are then presented.

2. Acoustic Model Development

This section describes experiments aimed at assessing the performance of a number of different acoustic models. Results are reported on an episode of NPR Marketplace recorded on 12 July 1996 (this episode in denoted as k960712 in the Hub-4E development test set). The language model used for the experiments on acoustic modelling was developed for the 1996 CU-CON system. Training data is the 132 million words of broadcast news texts, plus the 1995 Hub-4 texts which contain 108 million words and covers general North American business news. A trigram language model and a 65,532 word vocabulary were used for all the experiments.

2.1. MLP Acoustic Modelling

The MLP models used are fully connected with a single hidden layer consisting of 4000 logistic sigmoid units, and an output layer with softmax units. A cross-entropy error criterion is used during training, and this ensures that the model outputs are estimates of the *a posteriori* probability of phone class given the acoustic data [3]. The input to the network consists of nine contiguous frames of 12^{th} order perceptual linear prediction (PLP) coefficients plus log energy. The networks are trained using back-propagation and gradient descent. The gradient descent learning rate is adapted during training based on the cross-validation error. Learning proceeds with the initial (empirically set) learning rate. When the decrease in cross-validation error falls below a threshold the learning rate is reduced by a factor of two. This continues after each iteration. When the decrease in cross-validation error again falls below a threshold the learning rate is set to zero and training is stopped [4].

We looked at both gender-independent and gender-dependent acoustic modelling using MLPs. The mark up of the acoustic training data includes gender tags, and these were used to

Focus Condition	Gender Ind. Model	Gender Dep. Models
F0	24.0	25.3
F1	37.8	42.3
F2	38.2	43.6
F3	40.4	44.2
F4	38.5	41.8
F5	34.9	42.7
FX	65.0	66.3
OVERALL	32.7	35.5

Table 1: Word error rates by focus conditions for both gender independent and gender dependent MLP acoustic models.

produce training sets for male and female speakers. The selection of the gender at recognition time was based on the log likelihood of the decoded utterances. All the test data was decoded using both the male and female acoustic models, and the decoded utterance with the highest log likelihood selected to form the final system output. The results are shown in Table 1, and as can be seen the gender independent system performs better than the gender dependent system. This may be due to the relatively small (30%) proportion of training data from female speakers.

2.2. RNN Acoustic Modelling

In this section we report results for both context-independent and context-dependent recurrent neural network (RNN) acoustic models [5]. The first set of experiments examine the effect of both the size of the acoustic model and the amount of training data. Table 2 shows results for a model with 256 state units (83700 parameters) trained on 35 hours of data (denoted Model 1), and a model with 384 state units (174324 parameters) trained on 60 hours of data (denoted Model 2). It can be seen that increasing the model size and the training data results in an 8.2% relative reduction in word error rate.

Focus Condition	Model 1 (84k params)	Model 2 (174k params)
F0	25.4	22.5
F1	41.8	38.4
F2	38.2	43.6
F3	44.7	39.2
F4	38.2	32.1
F5	31.8	33.3
FX	61.8	63.4
OVERALL	34.3	31.5

Table 2: Word error rates by focus conditions for RNN acoustic models with different numbers of parameters.

Comparing the results from Tables 1 and 2 shows that there is little difference in performance between the gender-independent MLP system, and a system using an RNN acoustic model (Model 2). Indeed, the performance difference between the two systems is not significant at $p < 0.05$[1]. However, the MLP acoustic model has four times the number of parameters of the RNN model.

Focus Condition	CI System	Number of CD phone models			
		589	697	792	1002
F0	22.5	20.1	19.9	20.5	21.2
F1	38.4	34.6	33.7	35.5	34.5
F2	43.6	45.5	40.0	39.1	43.6
F3	39.2	32.2	31.4	28.8	31.2
F4	32.1	30.9	31.2	29.7	29.4
F5	33.3	35.4	34.4	34.9	37.5
FX	63.4	63.8	60.6	61.0	63.4
OVERALL	31.5	28.9	28.2	28.5	29.2

Table 3: Word error rates by focus conditions for different numbers of context-dependent phone models.

We have also investigated the effect on word error rate of the number of word-internal context-dependent phone models used. A brief description of context-dependent phone modelling in the ABBOT system is given in Section 4.3. Table 3 shows results for systems with different numbers of context-dependent phone models. It can be seen that the number of context-dependent models has only a small effect on recognition performance. The differences between each of the context-dependent systems are not significant at $p < 0.05$. However, introducing context-dependent models provides a significant (at $p < 0.05$) improvement over a context-independent system.

3. Incorporating Syllable Boundary Information

This section reports experiments aimed at improving recognition accuracy by incorporating syllable boundary information during search. Previous research on detecting syllable boundaries and using this information to improve recognition accuracy has been reported [7, 8]. In this work we use the method of Wu *et al* [7].

3.1. Detecting Syllable Boundaries

The broadcast news training data does not include syllable boundary or phonetic alignment information. An automatic procedure for determining syllable boundaries is therefore required. The method used in this work is based on deriving syllable boundaries from phonetic alignments. The first

[1] Significance tests were performed using the two-tailed matched pairs method described in [6]

step in determining the syllable boundaries is to produce pronunciations with tagged syllable boundaries. Syllable tagged pronunciations are required for every word in the training data. This was done automatically using the NIST software *tsylb2*[2]. The first phone of each syllable is tagged as an onset phone. Viterbi forced alignment is then used to determine phone alignments for the training data. These can be used in conjunction with the syllable tagged lexicon to derive the syllable onsets.

A single hidden layer, fully connected MLP with 500 hidden units was trained to estimate the probability that a given frame is a syllable onset. The input to this MLP consists if 9 contiguous frames of PLP features. For the purposes of training, the syllable onsets were represented as a series of four frames, with the initial frame corresponding to the actual onset derived from the phonetic alignments.

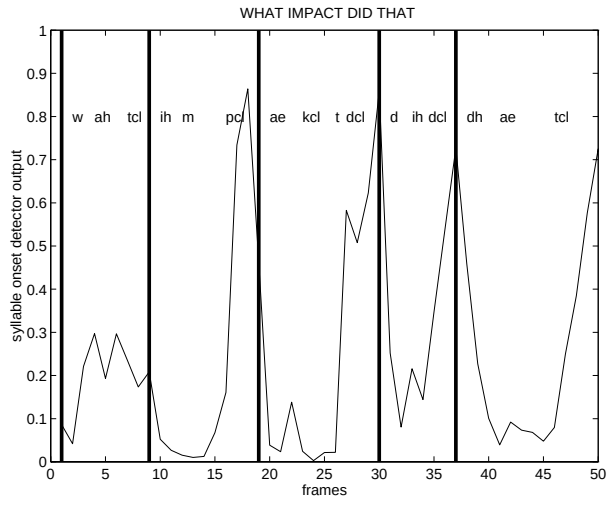

Figure 1: Example of the output of the syllable onset detector for the utterance "what impact did that". The vertical lines are the syllable onsets as derived from Viterbi aligned phone labels.

A simple numeric threshold applied to the probability estimates generated by the neural network determined the identification of any frame as a syllable onset. This method correctly detected 92% of the onsets derived from the phonetic alignments. However, this method also detected syllable onsets where there were none in 30% of frames outside the four-frame window defined for training. This effect can be seen in Figure 1 which shows an example of the neural network output. The width of the onsets detected tends to be much wider than the four-frame window used during training.

[2]The actual syllabification of the lexicon was done by Eric Fosler, of the International Computer Science Institute.

3.2. Syllable based Decoding

The NOWAY [9, 10] stack decoder was used to incorporate syllable boundary information in the decoding process. The context-independent phones may occur both at a syllable onset, or not directly after the syllable onset. This can be seen in the example pronunciation shown below in which the schwa (ax) occurs both at the beginning of the first syllable, and as the second phone of the last syllable. Phones that occur at syllable onsets are tagged with _on.

ABATEMENTS = { ax_on bcl b_on ey tcl m_on ax n tcl s }

Therefore two phone models are required for each context independent phone in the system, one model for when the phone occurs at a syllable onset, and one when it does not. The same acoustic model is used to generate the observation probabilities for the syllable onset phones and the standard (i.e.. not at syllable onsets) phones. This assumes that the realisation of any particular phone is not affected by whether or not it is the onset of a syllable. The observation probabilities of the onset phone models are set to zero when no onset is detected, and to those of the standard model when a syllable onset is detected. This effectively means that the decoder can only choose syllable onset phones when a syllable onset is detected, and thus allows the incorporation of syllable boundary information into a standard decoder.

Focus Condition	Standard CI system	CI + syllable onset system
F0	22.5	21.1
F1	38.4	32.1
F2	43.6	40.0
F3	39.2	37.1
F4	32.1	31.5
F5	33.3	32.3
FX	63.4	59.3
OVERALL	31.5	28.8

Table 4: Word error rates by focus conditions for a context-independent system, and a context-independent system incorporating syllable boundary information.

The results for context-independent systems with and without syllable boundary information can be seen in Table 4. Incorporating syllable onset information has reduced the word error rate for each of the focus conditions, and resulted in an overall reduction in word error rate of 8.6% (which is significant at $p < 0.05$).

4. The CU-CON Evaluation System

This section describes the CU-CON system used for the 1997 Hub-4E evaluation. This includes a description of the audio

segmentation, the front-end, the acoustic and language models, and the recognition procedure. The results on the development data presented in Sections 2 and 3 led to a number of decisions when designing the evaluation system. Firstly, although the performance of MLP and RNN acoustic models is similar, RNN models are more compact and context-dependent models are quick to construct. We therefore decided to use RNN acoustic models. Secondly, it was decided not to use syllable boundary information because initial experiments with a context-dependent system suggested that little or no gain was achieved.

4.1. Audio Segmentation

The CU-CON system used tools provided by NIST to perform audio segmentation. These tools implement the method of Siegler et al. [11]. Means and variances are estimated for a two second window placed at each point in the audio stream. A Kullback-Liebler distance between successive windows is then computed, and when this reaches a local maximum a new segment boundary is marked. The tools also classify the segments as either full or telephone bandwidth. This is accomplished by building Gaussian mixture models for each of the segments. Maximum likelihood selection of the class given the segment is performed by comparing the segment models to models trained on known bandwidth data. The tools also perform clustering of the segments, but this was not used by the CU-CON system.

4.2. Acoustic Feature Representation

Two sets of acoustic features are used by the 1997 CU-CON system: MEL+, a 20 channel mel-scaled filter bank with energy, degree of voicing, and pitch [12], and PLP, 12th order cepstral coefficients derived using perceptual linear prediction and log energy [13]. The MEL+ and PLP features were computed from 32 msec windows of the speech waveform every 16 msec. To increase the robustness of the system to environmental conditions, the statistics of each feature channel were normalised to zero mean with unit variance over each segment.

4.3. Acoustic Modelling

The basic acoustic modelling system is a recurrent neural network. The network maintains an internal state which provides a mechanism for modelling acoustic context and the dynamics of the acoustic signal. The output vector produced by the RNN acoustic model is an estimate of the *a posteriori* probability of each of the phone classes, i.e.

$$y_i(t) \simeq \Pr(q_i(t)|\mathbf{u}_1^{t+4}) \quad (1)$$

where $q_i(t)$ is state i at time t and $\mathbf{u}_1^t = \{\mathbf{u}(1), \ldots, \mathbf{u}(t)\}$ is the input from time 1 to t. The training approach is based on Viterbi training. Each frame of training data is assigned a phone label based on an utterance orthography and the current model. The recurrent network is then trained – using the back-propagation-through-time algorithm [14] – to map the input acoustic vector sequence to the phone label sequence. The labels are then reassigned and the process iterates. A more detailed description of the RNN architecture and the training algorithm can be found in [5, 1].

As shown in Section 2.2 context-dependent phone models lead to improved performance, and the CU-CON system uses word-internal CD models. The context-dependent phone models are chosen using a decision tree. The decision tree is constructed using rules that are based on the left and right contexts. A tree is grown for each monophone in the system [15]. This allows for sufficient statistics for training and keeps the system compact (allowing fast context training). The method used to implement CD phone models is based on the factorisation of conditional context-class probabilities [16]. The joint *a posteriori* probability of context class j and phone class i is given by

$$y_{ij}(t) = y_i(t) y_{j|i}(t). \quad (2)$$

The RNN estimates $y_i(t)$, and single-layer networks or "modules" are used to estimate the conditional context-class posterior, $y_{j|i}(t)$. The input to each module is the internal state of the recurrent network, since it is assumed that the state vector contains all the relevant contextual information necessary to discriminate between different context classes of the same monophone.

4.4. Acoustic Model Training

Different acoustic model sets were used for wide-band and telephone bandwidth data. A total of eight different acoustic models have been used. Four acoustic models have been used for wide-band data. All the wide-band models use PLP acoustic features, and estimate 697 word-internal context-dependent phone probabilities. Training data is from the 104 hours of broadcast news data. In total this contains approximately 76 hours of transcribed data. The average log likelihood per frame was computed during Viterbi alignment and those segments with poor scores were not used for training. This resulted in a total training set containing 60 hours of data.

Model	Parameters	Data Segments	Order
WB-1	174k	All	forward
WB-2	174k	All	backward
WB-3	84k	F0+F1	forward
WB-4	84k	F0+F1	backward

Table 5: Training data and model size for the wide-band acoustic models.

The conditions for training the wide-band models are shown in Table 5. Models were trained with the data presented both

forward and backward in time. This produces different acoustic models due to the fact that the RNN is time-asymmetric. Models were also produced using just the baseline planned studio speech (F0) and the spontaneous studio speech (F1) (about 31 hours of data). It has been shown that a significant performance improvement is achieved by merging multiple recurrent networks [17]. The output of the four wide-band models is merged in the log domain, i.e.

$$\log y_i(t) = \frac{1}{K} \sum_{k=1}^{K} \log y_i^{(k)}(t) - Z \qquad (3)$$

where Z is a constant chosen to insure that $\mathbf{y}$ is a valid probability distribution.

Four acoustic models have also been used for telephone bandwidth data. Each of the models estimate 604 word-internal context-dependent phone probabilities. The training data was taken from the 50 hours of broadcast news data released for the 1996 Hub-4 evaluation. All the models have a total of approximately 84k parameters. Forward and backward models were trained for both PLP and MEL+ features. All models were adapted to telephone bandwidth by mean of a linear input network (LIN). LIN adaptation has been successfully applied to connectionist HMM systems for supervised speaker adaptation [18], unsupervised speaker adaptation [19], and unsupervised channel adaptation [20, 21]. A linear mapping is created to transform the acoustic vector, and during recognition this transformed vector is fed as input to the RNN.

The F2 data was marked as either having low or medium fidelity. As with the 1996 evaluation we reclassified all the F2 data into narrow or wide band data based on the power in the upper 4kHz of the spectrum [22]. However, merely averaging the power in the upper 4kHz of a segment would bias the classification due to the relative number of voiced and unvoiced sections in a segment. To account for this we multiplied the energy in the upper 4kHz of each frame by the estimated probability of the frame representing an unvoiced segment. We chose a threshold for the choice of narrow bandwidth and full bandwidth by manually classifying a small proportion of the F2 segments. After setting this threshold all the F2 segments were relabelled. A LIN was trained for each model on the narrow bandwidth F2 data. These adapted models were used on the evaluation data classified as narrow bandwidth. As with the wide-band models the outputs of the telephone bandwidth models were merged in the log domain.

4.5. Language Model and Lexicon

4-gram and trigram backoff language models were trained from the LDC broadcast news training texts, the transcriptions of the broadcast news training data, the 1995 non-financial newswire (H4) texts, the 1995 financial newswire (H3) texts, and the 1995 Marketplace training data transcriptions. The language models were constructed using version 2.03 of the CMU-Cambridge SLM Toolkit. The Witten-Bell discounting method was used for both the 4-gram and trigram models. The language models contained 7.0 million bigrams, 24.1 million trigrams, and 4.7 million 4-grams.

The recognition lexicon contains 65,532 words, and was developed for the 1996 Hub-4 evaluation. Pronunciations for training were generated using a technique in which known pronunciations (from the LIMSI 1993 WSJ Lexicon) are segmented to form letter to sound rules which are then used to produce pronunciations for new words.

4.6. Recognition Procedure

In contrast to previous ABBOT systems the 1997 CU-CON system uses a two-pass recognition procedure. The first pass uses a trigram language model and is used to produce lattices. The NOWAY stack decoder is used for this first pass. A stack based lattice to n-best decoder is then used to produce 1-best hypotheses from the lattices. A 4-gram language model is used for this second pass. Note that no test set adaptation is performed during either recognition pass.

4.7. Results

This section presents results on the 1997 Hub-4E evaluation data. Table 6 shows the error rate of the system with a trigram language model (the first pass), and a 4-gram language model. As can be seen the use of a 4-gram language model has reduced the overall word error rate by 3.2%, which is significant at $p < 0.001$. The perplexity of the trigram language model is 179, and 166 for the 4-gram model. The overall out-of-vocabulary rate was 1.16%.

	3-gram LM	**4-gram LM**
Substitutions	17.9	17.4
Deletions	5.9	5.6
Insertions	4.2	4.1
OVERALL WER	27.9	27.1

Table 6: Word error rates for the CU-CON evaluation system with trigram and 4-gram language models.

The word error rate by focus condition is shown in Table 7. The use of a 4-gram language model has reduced the error rate for each of the focus conditions.

5. Future Work

We have examined the hypotheses produced by both context-independent and context-dependent systems on the development data on a per-segment basis. Initial results suggest that although the overall error rate of the CD system is significantly lower than the CI system, there are a relatively large number of segments for which the CI system produces the

Focus	N$^{o.}$ Words	Word Error Rate	
		3-gram LM	4-gram LM
F0	13197	15.9	15.5
F1	6566	27.3	26.3
F2	4882	38.9	37.5
F3	1571	36.7	35.1
F4	3350	32.6	31.2
F5	669	25.3	25.7
FX	2599	59.6	59.1
OVERALL	32834	27.9	27.1

Table 7: Word error rates by focus condition.

most accurate hypothesis. This suggests that if the hypotheses of the two systems could be combined in a suitable manner the overall error rate could be reduced. We are currently working on confidence based methods for combining the system outputs.

6. Acknowledgements

This work was partially funded by ESPRIT project 20077 SPRACH. The work on incorporating syllable boundary information is based on ideas from Steve Greenberg, Nelson Moragn, Su-Lin Wu, Mike Shire, and Eric Fosler from the International Computer Science Institute [7]. The authors wish to thank Eric Fosler and Su-Lin Wu for the assistance in implementing the syllable boundary experiments.

References

1. A.J. Robinson, M.M. Hochberg, and S.J. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 19. Kluwer Academic Publishers, 1995.
2. Nelson Morgan and Hervé Bourlard. Continuous Speech Recognition. *IEEE Signal Processing Magazine*, 12(3):24–42, May 1995.
3. M.D. Richard and R.P. Lippmann. Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation*, (3):461–483, 1991.
4. N. Morgan and H. Bourlard. Generalization and Parameter Estimation in Feedforward Nets: Some Experiments. In D.S. Touretzsky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.
5. A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.
6. L. Gillick and S.J. Cox. Some Statistical issues in the Comparison of Speech Recognition Algorithms. *International Conference on Acoustics, Speech and Signal Processing*, 1:532–535, 1989.
7. S-L. Wu, M.L. Shire, S. Greenberg, and N.Morgan. Integrating Syllable Boundary Information into Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing*, 2:987–990, April 1997. Berlin.
8. M.J. Hunt, M. Lennig, and P. Mermelstein. Experiments in Syllable-based Recognition of Continuous Speech. *International Conference on Acoustics, Speech, and Signal Processing*, 3:880–883, April 1980. Denver, Colorado.
9. S.J. Renals and M.M. Hochberg. Decoder Technology for Connectionist Large Vocabulary Speech Recognition. Technical Report CS-95-17, Dept. of Computer Science, University of Sheffield, 1995.
10. S. Renals and M. Hochberg. Efficient Evaluation of the LVCSR Search Space Using the NOWAY Decoder. *International Conference on Acoustics, Speech, and Signal Processing*, 1:149–152, 1996.
11. M.A. Siegler, U. Jain, B. Raj, and M. Stern. Automatic Segmentation, Classification and Clustering of Broadcast News. *DARPA Speech Recognition Workshop*, February 1997. Westfields International Conference Center, Chantilly, Virginia.
12. A.J. Robinson. Several Improvements to a Recurrent Error Propagation Network Phone Recognition System. Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department, September 1991.
13. H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.
14. P.J. Werbos. Backpropagation Through Time: What Does It Mean and How to Do It. In *IEEE*, volume 78, pages 1550–60, October 1990.
15. D.J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*. PhD thesis, Cambridge University Engineering Department, 1996.
16. D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge, MA 02142-1399, 1996.
17. M.M. Hochberg, G.D. Cook, S.J. Renals, and A.J. Robinson. Connectionist Model Combination for Large Vocabulary Speech Recognition. In *Neural Networks for Signal Processing*, volume IV, pages 269–278, 1994.
18. J. Neto, L. Almeida, M.M. Hochberg, C. Martins, L. Nunes, S.J. Renals, and A.J. Robinson. Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition Systems. In *Eurospeech*, pages 2171–2174, September 1995.
19. J.P. Neto, C.A. Martins, and L.B. Almeida. Unsupervised Speaker-Adaptation For Hybrid HMM-MLP Continuous Speech Recognition System. In *IEEE Speech Recognition Workshop*, pages 187–8, December 1995.
20. D.J. Kershaw, S. Renals, and A.J. Robinson. The 1995 ABBOT LVCSR System for Multiple Unknown Microphones. In *Int. Conf. in Spoken Language Processing*, October 1996.
21. D.J. Kershaw, A.J. Robinson, and S.J. Renals. The 1995 Hybrid Connectionist-HMM Large-Vocabulary Recognition System. In *ARPA Speech Recognition Workshop*, Harriman House, New York, February 1996.
22. G.D. Cook, D.J. Kershaw, J.D.M. Christie, and A.J. Robinson. Transcription of Broadcast Television and Radio News: The 1996 Abbot System. *DARPA Speech Recogniton Workshop*, February 1997. Westfields International Conference Center, Chantilly, Viginia.